Lecture 8: Summary Measures

Dipankar Bandyopadhyay, Ph.D.

BMTRY 711: Analysis of Categorical Data Spring 2011

Division of Biostatistics and Epidemiology

Medical University of South Carolina

Summary Measures of Association

Sometimes, individual comparisons are of interest.

Such as in our example of comparing the odds of a Fatal MI relative to No MI for the Aspirin group relative to placebo is valuable.

However, sometimes, a single summary measure is desirable.

Uncertainty Coefficient-Summary Measure for Nominal Categories

Theil (1970) proposed the index

$$U = -\frac{\sum_i \sum_j log(\pi_{ij}/\pi_{i}.\pi._{j})}{\sum_j \pi._{j} log \pi._{j}}$$

A value of $U = 0$ indicates independence of X and Y.

A value of $U = 1$ indicates that $\pi_{j|i} = 1$ for some j at each level of i.

The key limitation of this measure is that values of U are hard to interpret.

For example, if $U = .30$, is that a small or large association?

Example of Uncertainty Coefficient

Recall our myocardial infarction example.

We can calculate the joint probabilities as

Probabilities				
	Myocardial Infarction			
	Fatal Attack	Nonfatal Attack	No Attack	
Placebo	0.00081555	0.007747723	0.491368764	0.499932038
Aspirin	0.000226542	0.004485524	0.495355897	0.500067962
	0.001042091	0.012233247	0.986724661	1

Using the previous definition, it can be shown that U equals

$$U = -\frac{0.000625012}{-0.074212678} = 0.0084$$

Calculations in SAS

```
data uncert;
input i j count @@;
 cards;
 1 1 18  1 2 171  1 3 10845
 2 1 5  2 2 99  2 3 10933
;
run;
proc freq;
 tables i*j /measures;
 weight count;
run;
```

Selected Output

```
Statistic                                    Value       ASE
------------------------------------------------------------

Uncertainty Coefficient C|R                 0.0084    0.0031  <- our result
Uncertainty Coefficient R|C                 0.0009    0.0003
Uncertainty Coefficient Symmetric           0.0016    0.0006

Sample Size = 22071
```

Interpretation?

Ordinal Trends

Although the interpretation of U is difficult, when X and Y are both ordinal, there are additional measures to consider.

Monotone Trends:

1. Monotonically Increasing: As levels of X increase, the levels of the response, Y, increase
2. Monotonically Decreasing: As levels of X increase, the levels of the response, Y, decrease

We want to develop a single measure, similar to a correlation, that summarizes these trends.

Definitions:

1. A pair of subjects is *Concordant* if the subject ranked higher on X and also ranks higher on Y
2. A pair of subjects is *Discordant* if the subject ranked higher on X but ranks lower on Y
3. The pair is *tied* if both rank the same on X and Y

- Denote,

$$C = \text{Total number of concordant pairs}$$

$$D = \text{ Total number of discordant pairs}$$

- Then, Gamma (Goodman and Kruskal 1954) is defined as

$$\gamma = \frac{C - D}{C + D}$$

- However, this calculation is a little more involved than first observation.
- Lets explore the calculation for a 2×2 table

		Columns (j)	
		1	2
Rows (i)	1	18	171
	2	5	99

- Lets begin by estimating the number of concordant “pairs”
- Recall, a concordant pair must be greater in X and Y or Less in X and Y
- For Cell (1,1), there are 99 observations (the cell 2,2). Note: For the rows, 2>1 and for the columns 2>1
- Since cell (1,1) has 18 observations, we have 18*99 concordant pairs related to cell (1,1) (SHOW Peas in a Pod illustration)
- Likewise, for cell (2,2) (note: the only cell in which k<2 and l<2 for some pair (k,l) is cell (1,1)), there are 18 observations
- Thus, we have 99*18 concordant pairs for Cell (2,2)
- In total, we have $2 \times 18 * 99 = 3564$ concordant pairs
- Likewise the discordant pairs, D, are $2 \times 5 \times 171 = 1710$ so,

$$\gamma = \frac{3564 - 1710}{3564 + 1710} = 0.3515$$

Notes about Gamma

- Gamma treats the variables is symmetrically - you do not need to specify a response
- Gamma ranges from $-1 \leq \gamma \leq 1$
- When the categories are reversed, the sign of Gamma switches
- $|\gamma| = 1$ implies a perfect linear association
- When X and Y are independent, $\gamma = 0$. However $\gamma = 0$ does not imply independence (only that the Probability of a concordant pair is the same as the probability of a discordant pair, i.e. $\Pi_c = \Pi_d$)
- The general calculation formula for γ is as follows:

$$\gamma = \frac{P - Q}{P + Q}$$

where . . .

$$P = \sum_i \sum_j n_{ij} A_{ij}$$

where

$$A_{ij} = \sum_{k>i} \sum_{l>j} n_{kl} + \sum_{k<i} \sum_{l<j} n_{kl}$$

and

$$Q = \sum_i \sum_j n_{ij} D_{ij}$$

where

$$D_{ij} = \sum_{k>i} \sum_{l<j} n_{kl} + \sum_{k<i} \sum_{l>j} n_{kl}$$

Example

Consider the following data

Cross-Classification of Job Satisfaction by Income				
	Job Satisfaction			
	Very Dissatisfied	Little Dissatisfied	Moderately Satisfied	Very Satisfied
< 15,000	1	3	10	6
15,000 - 25,000	2	3	10	7
25,000 - 40,000	1	6	14	12
> 40,000	0	1	9	11

We want to summarize how job satisfaction and income relate.

We could calculate γ by hand, but I think I'll opt for SAS

In SAS - Read in the Data

```
data test;
 input i j count;
 cards;
 1 1 1
 1 2 3
 1 3 10
 1 4 6
 2 1 2
 2 2 3
 2 3 10
 2 4 7
 3 1 1
 3 2 6
 3 3 14
 3 4 12
 4 1 0
 4 2 1
 4 3 9
 4 4 11
 ;
 run;
```

Summarize the Data

```
proc freq;
 tables i*j/measures;
 weight count;
 run;
```

Review Results

```
Statistics for Table of i by j

Statistic                              Value       ASE
------------------------------------------------------
Gamma                                 0.2211    0.1172 <--- Our result
Kendall's Tau-b                       0.1524    0.0818
Stuart's Tau-c                        0.1395    0.0753

Somers' D C|R                         0.1417    0.0764
Somers' D R|C                         0.1638    0.0878

Pearson Correlation                   0.1772    0.0907
Spearman Correlation                  0.1769    0.0955

Lambda Asymmetric C|R                 0.0377    0.0828
Lambda Asymmetric R|C                 0.0159    0.0273
Lambda Symmetric                      0.0259    0.0407

Uncertainty Coefficient C|R           0.0312    0.0197
Uncertainty Coefficient R|C           0.0258    0.0167
Uncertainty Coefficient Symmetric     0.0282    0.0181

Sample Size = 96
```

Summary of Gamma

$\widehat{\gamma} = 0.2211$ with SE =0.1172, so an approximately 95% confidence interval can be calculated as

$$CI_{95\%} = 0.2211 \pm 1.96(0.1172) = (-0.0086, 0.4508)$$

Therefore at the $\alpha = 0.05$ level, there is insufficient evidence to support the hypothesis that a linear trend exists in the data.

In other words, there is no evidence to support an association of job satisfaction and income.

Over the next few lectures, we will examine additional ways of summarizing $I \times J$ contingency tables.

Generalized Table

- Lets suppose that we have an $I \times J \times Z$ contingency table.
- That is, There are I rows, J columns and Z layers.

Conditional Independence

We want to explore the concept of conditional independence. But first, lets review some probability theory.

Recall, two variables A and B are independent if and only if

$$P(AB) = P(A) \times P(B)$$

Also recall that Bayes Law states for any two random variables

$$P(A|B) = \frac{P(AB)}{P(B)}$$

and thus, when X and Y are independent,

$$P(A|B) = \frac{P(A)P(B)}{P(B)} = P(A)$$

Conditional Independence

Definitions:

In layer k where $k \in \{1, 2, \ldots, Z\}$, X and Y are conditionally independent at level k of Z when

$$P(Y = j|X = i, Z = k) = P(Y = j|Z = k), \quad \forall i, j$$

If X and Y are conditionally independent at ALL levels of Z, then X and Y are CONDITIONALLY INDEPENDENT.

Application of the Multinomial

Suppose that a single multinomial applies to the entire three-way table with cell probabilities equal to

$$\pi_{ijk} = P(X = i, Y = j, Z = k)$$

Let

$$\begin{aligned} \pi_{\cdot jk} &= \sum_{X} P(X = i, Y = j, Z = k) \\ &= P(Y = j, Z = k) \end{aligned}$$

Then,

$$\pi_{ijk} = P(X = i, Z = k)P(Y = j | X = i, Z = k)$$

by application of Bayes law. (The event $(Y = j) = A$ and $(X = i, Z = k) = B$).

Then if X and Y are conditionally independent at level z of Z,

$$\begin{aligned}
\pi_{ijk} &= P(X=i, Z=k)P(Y=j|X=i, Z=k) \\
&= \pi_{i\cdot k}P(Y=j|Z=k) \\
&= \pi_{i\cdot k}P(Y=j, Z=k)/P(Z=k) \\
&= \pi_{i\cdot k}\pi_{\cdot jk}/\pi_{\cdot\cdot k}
\end{aligned}$$

for all i, j, and k.

Example

Suppose we look at the response (success, failure) (Y) for Treatments (A,B) (X) for a given center (1,2) (Z). There is a total sample size of $n = 100$

		Response	
Clinic	Treatment	Success	Failure
1	A	18	12
	B	12	8
2	A	2	8
	B	8	32
Total	A	20	20
	B	20	40

Recall the MLE for any parameter of the multinomial is n_{ijk}/n.

We want to examine whether or not the Response is independent of Treatment for each clinic.

Let π_{111} be the response probability for a Success of Treatment A at Clinic 1.
Then,

$$\pi_{111} = 18/100 = .18$$

Using the definition of conditional independence, X and Y are conditionally independent if and only if

$$\pi_{ijk} = \pi_{i \cdot k} \pi_{\cdot jk} / \pi_{\cdot \cdot k}, \quad \forall i, j, k$$

Then,

$$\pi_{1 \cdot 1} = (18 + 12)/100 = .30$$

$$\pi_{\cdot 11} = (18 + 12)/100 = .30$$

$$\pi_{\cdot \cdot 1} = (18 + 12 + 12 + 8)/100 = .50$$

Thus,

$$\begin{aligned} \pi_{1 \cdot 1} \pi_{\cdot 11} / \pi_{\cdot \cdot 1} &= (.3)(.3)/.5 \\ &= 9/50 \\ &= .18 \end{aligned}$$

So for $\{X = 1, Y = 1, Z = 1\}$ X and Y are conditionally independent.

We need to verify the conditional independence holds for other combinations of i, j, k.

For (212) (i.e., A success for treatment B at Site 2)

$$\pi_{212} = 8/100 = .08$$

$$\pi_{2\cdot 2} = (8 + 32)/100 = .40$$

$$\pi_{\cdot 12} = (2 + 8)/100 = .10$$

$$\pi_{\cdot\cdot 2} = (2 + 8 + 8 + 32)/100 = .50$$

Thus,

$$\begin{aligned} \pi_{2\cdot 2}\pi_{\cdot 12}/\pi_{\cdot\cdot 2} &= (.4)(.1)/.5 \\ &= 4/50 \\ &= .08 \end{aligned}$$

There are other combinations to verify; however, we will stop here and say that X and Y are conditionally independent given Z

Conditional Independence and Marginal Independence

We have just shown that the treatment and response are conditionally independent given a clinic.

Does this imply that treatment and response are independent in general?

That is, does

$$\pi_{ij\cdot} = \pi_{i\cdot\cdot}\pi_{\cdot j\cdot} \quad ?$$

According to the definition of conditional independence,

$$\pi_{ijk} = \pi_{i\cdot k}\pi_{\cdot jk}/\pi_{\cdot\cdot k}, \quad \forall i, j, k$$

and since $\pi_{ij\cdot} = \sum_k \pi_{ijk}$,

$$\sum_k \pi_{ijk} = \sum_k \pi_{i\cdot k}\pi_{\cdot jk}/\pi_{\cdot\cdot k}$$

Since the three probabilities on the right hand side of the equation all involve k, no simplification can be made.
Thus,

$$\sum_k \pi_{ijk} \neq \pi_{i\cdot\cdot}\pi_{\cdot j\cdot}$$

That is, CONDITIONAL INDEPENDENCE does not imply MARGINAL INDEPENDENCE.

We were interested in Conditional Associations.

- For a partial table $z \in Z$, the association of $OR_{XY(z)}$ is called a Conditional Odds Ratio
- X and Y are conditionally independent if $OR_{XY(z)} = 1 \quad \forall z \in Z$

From our example

$$OR_{\text{Site 1}} = \frac{18 \times 8}{12 \times 12} = 1$$

and

$$OR_{\text{Site 2}} = \frac{2 \times 32}{8 \times 8} = 1$$

The marginal association of X and Y is

$$OR = \frac{20 \times 40}{20 \times 20} = 2$$

Therefore, since $OR_{(1)} = OR_{(2)} = 1$, X and Y are conditionally independent given Z (or center) where as X and Y are NOT INDEPENDENT.

Also, this example illustrates a homogeneous XY association since

$$OR_{(1)} = OR_{(2)}$$

Also note, it is much easier to use the fact that $OR = 1$ instead of the probability statements to show independence, but how do you prove this?

Proof:
Let $OR_{(k)} = \pi_{11k}\pi_{22k}/\pi_{12k}\pi_{21k}$ be the Odds Ratio for the k^{th} partial table.

If X and Y are conditionally independent at level k of Z then,

$$
\begin{aligned}
OR_{(k)} &= \pi_{11k}\pi_{22k}/\pi_{12k}\pi_{21k} \\
&= \frac{\left(\frac{\pi_{1\cdot k}\pi_{\cdot 1k}}{\pi_{\cdot\cdot k}}\right)\left(\frac{\pi_{2\cdot k}\pi_{\cdot 2k}}{\pi_{\cdot\cdot k}}\right)}{\left(\frac{\pi_{1\cdot k}\pi_{\cdot 2k}}{\pi_{\cdot\cdot k}}\right)\left(\frac{\pi_{2\cdot k}\pi_{\cdot 1k}}{\pi_{\cdot\cdot k}}\right)} \\
&= 1
\end{aligned}
$$

Extensions to more than 2 dimensions

Suppose we want to study the effect of X on Y.

- For valid comparisons, we should control for factors that may be related to both X and Y.
- Those factors that are related to both are called confounding variables.
- Example
  Suppose we are interested in exploring the relationship of the death verdict on racial factors. The data we have available summarizes death penalty by the victim's race and the defendant's race.

Victims Race	Defendants Race	Death Yes	Penalty No	Percent Yes
White	White	53	414	11.3
	Black	11	37	22.9
Black	White	0	16	0.0
	Black	4	139	2.8
Total	White	53	430	11.0
	Black	15	176	7.9

Partial Tables

To control for a confounding variable Z, we need to look at the association of X on Y at a level of Z, $Z = 1, \ldots, z$.

- The z subtables are called partial tables
- Summing over Z (i.e., ignoring the effects of Z) results in a MARGINAL table.

In our example, we are going to control for the VICTIM'S RACE.

Conditional Associations

- For a partial table $z \in Z$, the association of X on Y is called a Conditional association
- Let $OR_{XY(z)}$ be defined as the Odds Ratio for partial table $z \in Z$.
- A table has homogeneous XY association when

$$OR_{XY(1)} = OR_{XY(2)} = \cdots = OR_{XY(Z)}$$

- However, if some of these associations are not equal, then the factor Z is described as an effect modifier.
- Think of an effect modifier as an interaction term - The conditional association of X on Y is dependent on the value of Z.

Example

Recall from the previous example,

We wish to study the effects of racial characteristics on whether persons convicted of homicide received the death penalty. Initially, lets looks at the 674 subjects classified by the Defendant's Race and Death Penalty

		Death Penalty		
		1	2	
Defendant's Race	1	53	430	483
	2	15	176	191
		68	606	674

Note that this table has been "collapsed" over victim's race.

The observed association (as measured by OR) of the defendant's race and death penalty is

$$OR = \frac{53 \cdot 176}{15 \cdot 430} = 1.45$$

White Victim's

If we evaluated only White Victim's, we would observe

		Death Penalty		
		1	2	
Defendant's Race	1	53	414	467
	2	11	37	48
		64	451	515

The observed OR of the defendant's race and death penalty for WHITE VICTIMS is

$$OR_{(white\ victims)} = \frac{53 \cdot 37}{11 \cdot 414} = 0.4306$$

black Victim's

If we evaluated only Black Victim's, we would observe

		Death Penalty		
		1	2	
Defendant's Race	1	0	16	16
	2	4	139	143
		4	155	159

The observed OR of the defendant's race and death penalty for BLACK VICTIMS is

$$OR_{(black\ victims)} = \frac{0 \cdot 139}{4 \cdot 16} = 0$$

Or in terms of the empirical logit

$$OR^{E}_{(black\ victims)} = \frac{(0+0.5)\cdot(139+0.5)}{(4+0.5)\cdot(16+0.5)} = 0.939$$

Simpson’s Paradox

- Sometimes the marginal association is in the opposite direction from the conditional associations.
- This is Simpson’s Paradox
- Our example illustrates the paradox
- Simpson’s Paradox is often one of the arguments when investigators try to draw causal effects from associations of X with Y.
- Another case of Simpson’s paradox is when there is a change in the magnitude of association
- Consider the following example

Example

Gender	Aortic Stenosis	Smoker Yes	No	Total
Males	Yes	37	25	62
	No	24	20	44
Females	Yes	14	29	43
	No	19	47	66
Combined	Yes	51	54	105
	No	43	67	110

- We want to study the association of smoking on aortic stenosis (narrowing of the aorta)
- We have stratified our sample based on gender (Males have higher risk of cardiovascular disease)
- We can use SAS to assist in the calculations

```
options nocenter;
data one;
 input gender aortic smoker count;
 cards;
1 1 1 37
1 1 2 25
1 2 1 24
1 2 2 20
2 1 1 14
2 1 2 29
2 2 1 19
2 2 2 47
;
run;
title "Partial Table: Males";
proc freq data=one;
 where gender = 1;
 tables aortic * smoker /chisq;
 weight count;
run;
title "Partial Table: Females";
proc freq data=one;
 where gender = 2;
 tables aortic * smoker /chisq;
 weight count;
run;
title "Marginal Table: Gender combined";
```

```
proc freq data=one;
 tables aortic * smoker /chisq;
 weight count;
run;
```

Selected Results

```
Statistics for Table of aortic by smoker for MALES *****

Statistic                     DF       Value      Prob
------------------------------------------------------
Chi-Square                     1      0.2774    0.5984

Statistics for Table of aortic by smoker for FEMALES *****

Statistic                     DF       Value      Prob
------------------------------------------------------
Chi-Square                     1      0.1753    0.6754

Statistics for Table of aortic by smoker COMBINED *****

Statistic                     DF       Value      Prob
------------------------------------------------------
Chi-Square                     1      1.9623    0.1613
```

Although the combined table isn't statistically significant, there is a change in the evidence for an association. This too is Simpson's paradox.